A Experimental Setup

A.1 Dataset Statistics

The SST dataset consists of 67,349 training, 872 validation, and 1,821 test samples with binary sentiment annotations. The AG News contains 120,000 training and 7,600 test samples with 4 classes.

A.2 Detailed Setup

We select model architectures to achieve a reasonable tradeoff (Tsipras et al., 2019) between nominal accuracy and robust accuracy using the validation set. In the SST word-level experiments, we use a 1-layer convolutional network with 100 kernels of width 5, followed by a ReLU, an average pool, and a linear layer. We use pre-trained 300-dimensional GloVe embeddings (Pennington et al., 2014), and use counter-fitted embeddings (Mrkšić et al., 2016) in Section 4.6. The pre-trained word embeddings are fixed during training. In the SST character-level experiments, we use a 1-layer convolutional network with 100 kernels of width 5, followed by a ReLU, an average pool, followed by a linear layer. We set the character embedding dimension to 150, randomly initialise them, and fine-tune the embeddings during training. In the AG News character-level experiments, we follow the setup in Zhang et al. (2015) using lower-case letters only and truncate the character sequences to have at most 300 characters during training. We use a 1-layer convolutional network with 100 kernels of width 10, followed by a ReLU, an average pool, and two fully-connected layers with 100 hidden units, followed by a linear layer. We set the character embedding dimension to 150, randomly initialise them, and fine-tune the embeddings during training. Note since the proposed technique is efficient, we can scale up to deeper networks for better nominal accuracy at the cost of verified accuracy, as the bounds become looser.

We use Adam (Kingma and Ba, 2015) as our optimisation method, perform early stopping, and tune our hyperparameters (learning rate, loss ratio κ) on the validation set.

B Additional Experimental Results and Discussion

B.1 Ease of Verification (Computation of True Robustness)

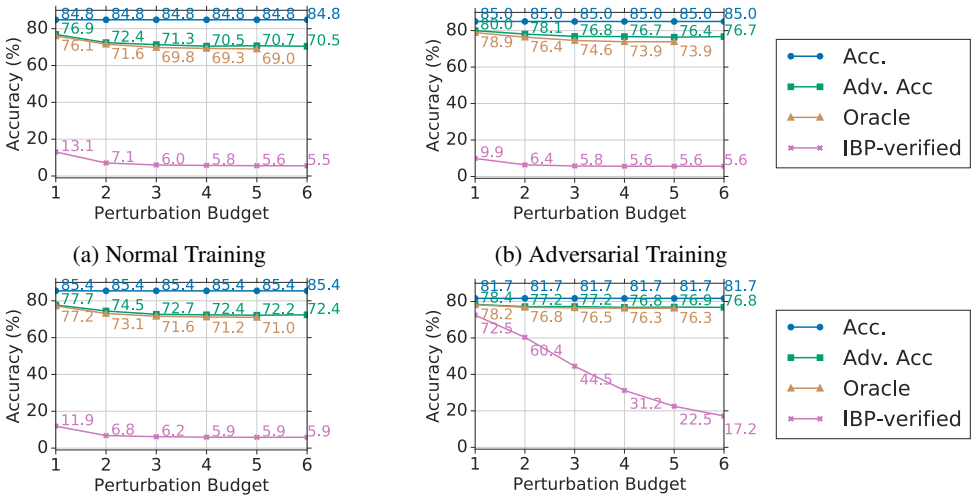
For every training method, we can compute the true robustness using exhaustive verification. However, this oracle is extremely computationally expensive (especially in character-level perturbations). On the other hand, verification via IBP provides a lower bound on the worst-case results, but this is generally loose for arbitrary networks. IBP-verifiable training succeeds in tightening these bounds and results in much improved rates of IBP-verification at test time, compared to all other training methods. We furthermore can observe that models trained to become verifiable (with IBP training objective) achieve better adversarial accuracy and exhaustively verified accuracy, with a small (or no) deterioration in nominal accuracy compared to normal training.

B.2 SST Word Embeddings Comparison

In Figures 5 and 6, we show the experimental results of different models and metrics using GloVe and counter-fitted embeddings, respectively.

B.3 AG News

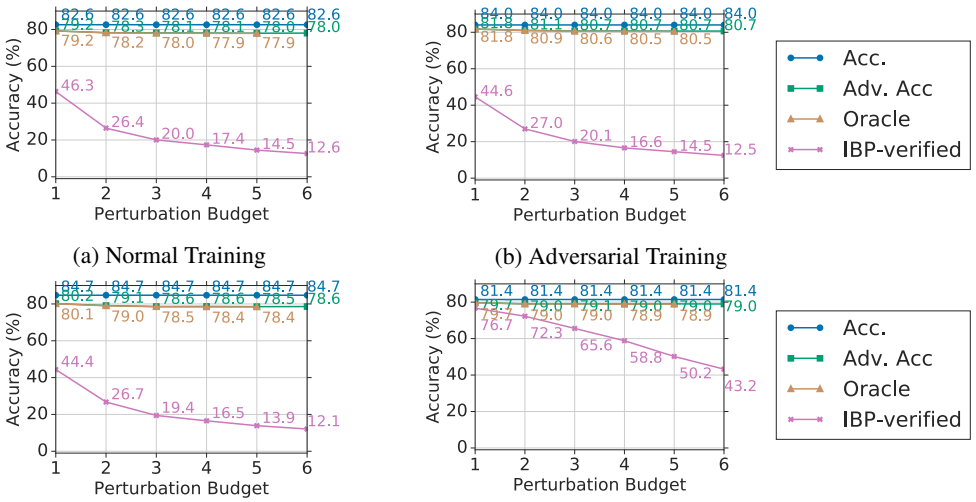
In Figure 7, we compare normal training, adversarial training, data augmentation, and verifiable training models with four metrics under various perturbation budgets on the AG News dataset at the character level. In Figure 7d, our verifiable trained model achieves not only the strongest adversarial and oracle accuracy, but achieves very tight bounds with respect to the oracle results. Note IBP verification only requires 2 forward passes to verify any examples, whereas oracle evaluation (exhaustive search) uses up to 260,282 forward passes for examining a single example at $\delta = 2$.



(a) Normal Training (b) Adversarial Training

(c) Data Aug. Training (d) Verifiable Training (IBP)

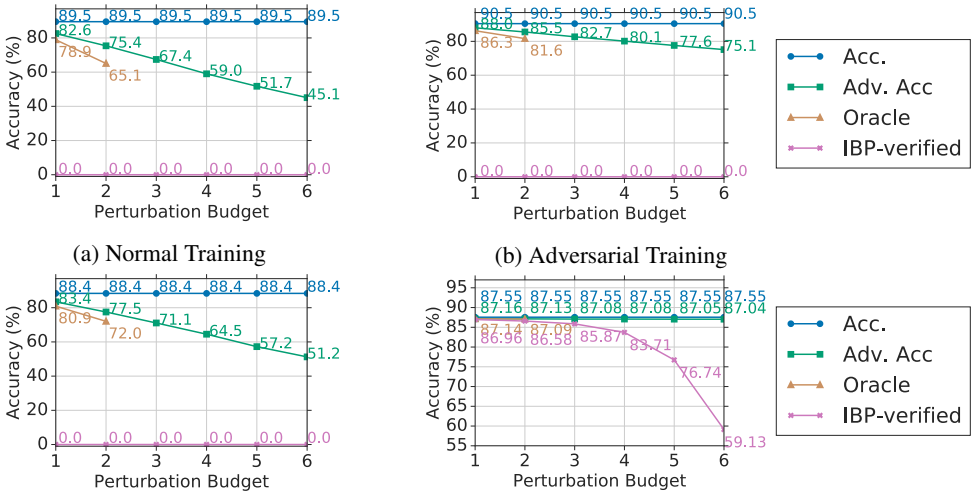
Figure 5: SST word-level models with different training objectives (trained at δ=3) using *GloVe* embeddings against different perturbation budgets in nominal accuracy, adversarial accuracy, exhaustively verified accuracy (Oracle), and IBP verified accuracy. Note that exhaustive verification is not scalable to perturbation budget 6 and beyond.



(a) Normal Training (b) Adversarial Training

(c) Data Aug. Training (d) Verifiable Training (IBP)

Figure 6: SST word-level models (trained at δ=3) using *counter-fitted* embeddings against different perturbation budgets in nominal accuracy, adversarial accuracy, exhaustively verified accuracy (Oracle), and IBP verified accuracy. Note that exhaustive verification is not scalable to perturbation budget 6 and beyond.



(a) Normal Training

(b) Adversarial Training

(c) Data Augmentation Training

(d) Verifiable Training (IBP)

Figure 7: AG News character-level models with different training objectives (trained at δ=3) against different perturbation budgets in nominal accuracy, adversarial accuracy, exhaustively verified accuracy (Oracle), and IBP verified accuracy. Note that exhaustive verification is not scalable to perturbation budget 3 and beyond.